An implicit representation of stimulus ambiguity in pupil size

Jackson E. Graves[a,1], Paul Egré[b,2], Daniel Pressnitzer[a,2], and Vincent de Gardelle[c,2]

[a]Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, 75005 Paris, France; [b]Institut Jean-Nicod, Département de philosophie & Département d'études cognitives, École normale supérieure, École des hautes études en sciences sociales, PSL University, CNRS, 75005 Paris, France; and [c]Centre d'Economie de la Sorbonne, Paris School of Economics & CNRS, 75013 Paris, France

Edited by Anna C. (Kia) Nobre, University of Oxford, Oxford, United Kingdom, and approved October 14, 2021 (received for review April 29, 2021)

To guide behavior, perceptual systems must operate on intrinsically ambiguous sensory input. Observers are usually able to acknowledge the uncertainty of their perception, but in some cases, they critically fail to do so. Here, we show that a physiological correlate of ambiguity can be found in pupil dilation even when the observer is not aware of such ambiguity. We used a well-known auditory ambiguous stimulus, known as the tritone paradox, which can induce the perception of an upward or downward pitch shift within the same individual. In two experiments, behavioral responses showed that listeners could not explicitly access the ambiguity in this stimulus, even though their responses varied from trial to trial. However, pupil dilation was larger for the more ambiguous cases. The ambiguity of the stimulus for each listener was indexed by the entropy of behavioral responses, and this entropy was also a significant predictor of pupil size. In particular, entropy explained additional variation in pupil size independent of the explicit judgment of confidence in the specific situation that we investigated, in which the two measures were decoupled. Our data thus suggest that stimulus ambiguity is implicitly represented in the brain even without explicit awareness of this ambiguity.

ambiguity | uncertainty | confidence | auditory perception | pupillometry

To infer the objects and characteristics of our environment, perceptual systems must deal with sensory data that is inherently incomplete and therefore ambiguous. The most common view to explain how we overcome this fundamental issue is that prior knowledge is combined with incoming inputs, taking the reliability of each source of information into account (see e.g., refs. 1 and 2). In this perspective, perception is thus described as a Bayesian inference process that estimates probabilities about the candidate object [(3), although see ref. 4 for a debate].

One central question in this framework is how the uncertainty is represented at the various stages of perceptual processing and, relatedly, how it may be explicitly reported by observers. Studies of metaperception have long shown that observers can have a good sense of the accuracy of their own perceptual decisions when they report their confidence (e.g., ref. 5). However, there are also clear cases of a disconnect between the objective uncertainty of perception as measured by performance and the subjective uncertainty reported by observers. Recently, stimuli such as #TheDress (6) in vision and Laurel/Yanny (7) in audition have provided vivid illustrations of a seemingly complete disconnect between ambiguous stimuli, as shown by the different perceptual interpretations reported by different observers and a very high confidence for all observers nonetheless. This pattern of high confidence, despite high variability in perceptual judgments, can also be demonstrated within the same observer. For instance, a naïve listener hearing pairs of tones with ambiguous frequency shifts perceives upward or downward shifts across different trials (8, 9), with high confidence in both cases (10).

In the present study, we ask whether this ambiguity, even though it is not consciously accessed by participants, might still be explicitly represented within the perceptual system. Specifically, we looked for a correlate of ambiguity in the pupil dilation signal. It is well established that pupil size fluctuates in response to cognitive processes, even when luminance is kept constant. Cognitive effects on pupil size have been found for various tasks such as mental arithmetic (11), memory [(12), see ref. 13 for a review], or control tasks (reviewed in ref. 14). Directly relevant to the question of ambiguity, pupil size is also affected by uncertainty in learning (15–19) and by uncertainty in perceptual decision tasks (12, 20, 21). At the neural level, it is thought that changes in pupil dilation reflect the activity of neuromodulatory networks, including norepinephrine (also called noradrenaline) and acetylcholine (22–24), which may for instance enhance sensory processing and cognitive flexibility (for reviews see e.g., refs. 25–27).

Perceptual uncertainty, which broadly includes cases of ambiguity, has been related to pupil dilation in two kinds of studies. Some studies have used stimuli approaching the discrimination threshold between two categories and measured larger pupil size for more difficult stimuli (20, 21, 28). In this case, decisions are typically made with low confidence, as sensory support for either response is weak, and participants are aware of the uncertainty (21). Another kind of study involves bistable perception, in which the uninterrupted presentation of an ambiguous stimulus triggers alternations between two clear percepts (for reviews see refs. 29 and 30). When participants have to monitor these alternations, pupil size was found to increase around the time of perceptual switches (31). Importantly, whether in threshold or bistable stimuli, pupil dilation

Significance

The perception of our environment usually comes with a sense of how uncertain or certain we are about what we perceive. In some cases, however, stimuli that are objectively ambiguous are not acknowledged as such, raising the question of whether uncertainty is not processed at all in these cases or whether instead, uncertainty signals might be computed but inaccessible to conscious awareness. Here, we show that the ambiguity of an auditory stimulus (measured through the variability of perceptual decisions over trials) is reflected in pupil dilation, despite participants being unaware of this ambiguity. This finding suggests that uncertainty signals in the brain are not always connected to conscious awareness.

Author Contributions: J.E.G., P.E., D.P., and V.d.G. designed research; J.E.G. performed research; J.E.G. analyzed data; and J.E.G., P.E., D.P., and V.d.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jackson.graves@ens.psl.eu.

[2]P.E., D.P., and V.d.G. contributed equally to this work.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2107997118/-/DCSupplemental.

Published November 24, 2021.

may reflect either uncertainty per se, or instead, it may reflect participants' explicit awareness of the uncertainty, the deliberate cognitive effort associated with resolving ambiguity, or even motor response preparation (e.g., ref. 32 for the case of bistability).

This distinction is critical: If pupil dilation indexes uncertainty per se, it reflects at least in part a fundamental and somewhat elusive ingredient of all probabilistic models of perception (33); otherwise, it could be a more generic consequence of arousal or cognitive effort. To test this distinction, we ask whether pupil dilation could be sensitive to stimulus uncertainty when participants are fully unaware of such uncertainty. This controls for the awareness of uncertainty and for conscious mental effort. We also ensured that task demands were identical for conditions of varying uncertainty, controlling for response-related effects.

In two experiments, we asked participants to report the perceived pitch change ("up" or "down") for pairs of so-called Shepard tones (8, 9). Shepard tones are created by stacking octave-related frequency components so that the dominant pitch–shift percept corresponds to the shortest log–frequency distance across components (8). Crucially, when the interval between two Shepard tones is a half-octave, there is no shortest log–frequency distance; the stimulus is intrinsically ambiguous, and the percept varies across listeners or even across repeated presentations in the same listener (8, 9). Here, the ambiguous nature of the stimuli was never mentioned to participants, whose behavioral responses indicated no explicit awareness of this ambiguity, confirming previous findings (10). By contrast, their pupil dilation was highest in the ambiguous conditions, and it was also correlated with the objective uncertainty in the stimuli (quantified via the entropy of perceptual decisions).

In sum, we present empirical evidence for a nonconscious representation of auditory uncertainty, that is reflected in pupil dilation in human listeners. This shows that the neuromodulatory systems involved in the conscious monitoring of uncertainty may also represent a form of unconscious uncertainty, consistent with this uncertainty being resolved before the perceptual content reaches awareness. Evidence for an early representation and resolution of uncertainty also echoes with the classic idea of perception as unconscious inference (see e.g., ref. 34). Finally, such a finding contributes to our understanding of the interplay between perceptual processing, perceptual awareness, and metacognition (e.g., ref. 35).

Results

Experiment 1: Response Variability and Pupil Dilation for Ambiguous Tone Shifts.

Behavioral evidence for unaccessed ambiguity. Participants (n = 20) were presented with pairs of Shepard tones with various intervals between the first and second tone while their pupil size was recorded. They indicated the direction of the change between the two tones through button presses after an enforced delay period to provide enough time to observe pupil size (Fig. 1*A*).

Fig. 1*B* illustrates the stimulus conditions. It is expected that cases for which there is clearly a shorter path in log– frequency space will provide unambiguous pitch–shift responses corresponding to this shortest path (e.g., an upward shift for a 2 semitone [ST] interval). In the 6 ST case, or half-octave, there is no shortest path, and responses are expected to vary. In the 0 ST case, there is no actual frequency shift, so responses are expected to be driven by internal noise and not stimulus ambiguity. We also introduced a last condition, termed ±2 ST. In this case, two nonambiguous frequency shifts were introduced in opposite directions. Here, as well, responses are expected to vary but for a third reason: the choice between two nonambiguous shifts. In order to allow participants to freely report their behavioral percept without forcing an inaccurate report for the 0 ST or 6 ST conditions, we instructed participants to choose between four response options: "up," "down," "both," or "neither."

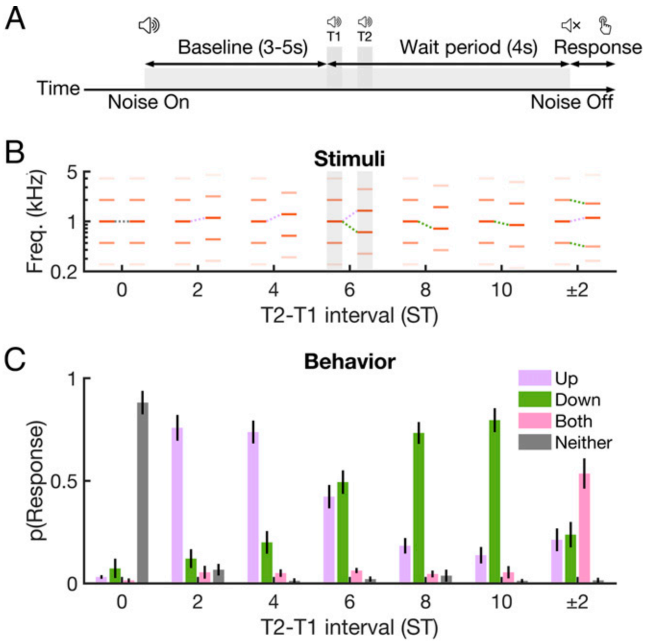

Fig. 1. Behavioral responses to ambiguous frequency shifts. (*A*) Schematic showing task design and timing on each trial. After a variable baseline period of 3 to 5 s, participants heard two tones (T1 and T2, 250 ms each) separated by 250 ms. At the end of the waiting period (4 s after T1 onset, cued by offset of background noise), participants made a keypress response categorizing the change between T1 and T2. (*B*) Schematic spectrograms showing stimuli in each condition of Experiment 1. Dashed lines indicate the shortest path between components, with upward and downward paths equally prominent at 6 and ±2 ST. (*C*) Probability of each behavioral response by condition in Experiment 1. Responses "up" and "down" are equivalently frequent at 6 and ±2 ST, but the response "both" is more frequent at ±2 ST. Error bars show ±1 SEM (n = 20).

Fig. 1*C* shows the behavioral responses of participants. As expected, in nonambiguous conditions, participants reported most often hearing a sound going up for shorter intervals (2 ST and 4 ST conditions) and down for larger intervals (8 ST and 10 ST conditions). In the 6 ST condition, however, they reported up and down equally often, demonstrating the ambiguity of this specific stimulus.

To test whether participants had explicit access to this ambiguity in Experiment 1, we gave them the opportunity to indicate whether they heard both directions or neither. They did not use these responses for the 6 ST condition, suggesting that they were unaware of the ambiguity there. We could verify nonetheless that these options were used appropriately otherwise. In the condition with 0 ST between the two Shepard tones (i.e., the first and second tone were identical), participants systematically reported hearing neither direction. In the condition where the stimulus contained both a +2 ST shift and a −2 ST shift, participants mostly heard both directions, as expected, but also sometimes up or down.

Pupil dilation is associated with response entropy and interval size. The ambiguity of the 6 ST condition is unaccounted for by explicit reports of participants, but by monitoring pupil size during the task, we could evaluate whether this ambiguity may have been represented at an implicit level. Fig. 2*A* illustrates for each condition the average pupil response evoked by the stimulus, which peaked at around 2 s after the stimulus onset and then slowly returned to baseline levels.

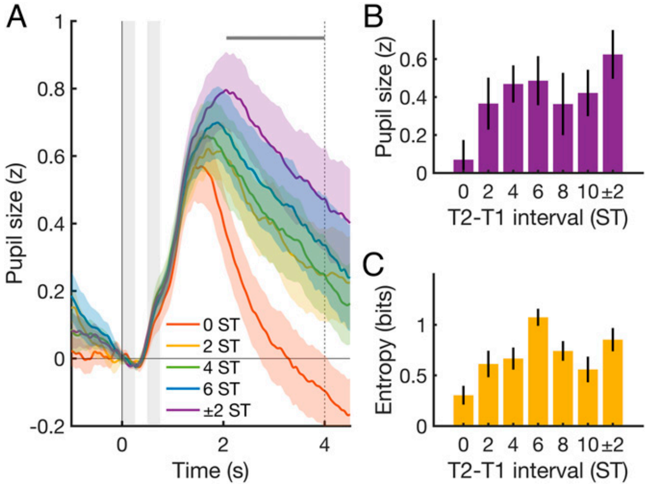

Fig. 2. Pupil dilation in Experiment 1. (*A*) Pupil dilation throughout the trial, baseline corrected from 100 ms before T1 onset to T2 onset, *z*-scored for each subject, and averaged across subjects in each condition. For plotting purposes, 2 and 10 ST are combined, and 4 and 8 ST are combined, as these conditions have equivalent interval sizes (note that they were separate conditions in our analyses). Gray regions indicate the presentation of T1 and T2. The dashed vertical line indicates the onset of the response window. The horizontal gray bar indicates the temporal cluster in which the main effect of condition on pupil size was significant. Colored regions show ±1 SEM ($n = 20$). (*B*) Pupil dilation (averaged over the identified cluster) for each interval size. (*C*) Entropy of behavioral responses for each interval size. Error bars show ±1 SEM ($n = 20$).

To evaluate whether this evoked pupil dilation varied across conditions, we conducted a one-way, repeated-measures ANOVA at each time point from 1 to 4 s after stimulus onset, and the resulting F statistics were used in a cluster-based permutation test. This procedure identified a significant cluster for the effect of condition between 2.06 and 4 s after stimulus onset (total F = 375.01, $P < 0.001$). Pairwise comparisons within this cluster (Bonferroni corrected $\alpha = 0.0024$ for 21 comparisons between seven conditions) revealed that the 6 ST, ±2 ST, and 4 ST conditions evoked greater pupil dilation than the 0 ST condition, with no other significant differences (see Fig. 2*B* for pupil size within this cluster).

Considering the two conditions with the greatest pupil dilation (6 and ±2 ST), the increased pupil dilation at ±2 ST is expected because not only did participants exhibit behavioral variability here but also some degree of explicit awareness of this variability (as evidenced by the occasional use of the both response). More interesting, and perhaps surprising, is the increased pupil dilation at 6 ST, because here participants showed no evidence of awareness of the inherent ambiguity in the stimulus. Participants nevertheless demonstrated behavioral variability for the 6 ST condition, so results for the ±2 ST and 6 ST conditions suggest that pupil dilation is associated with overall behavioral variability, with or without conscious awareness of such variability.

To test this hypothesis, we quantified behavioral variability by computing the entropy of responses over the course of the whole experiment. Looking at pupil dilation together with the entropy of behavioral responses, our data suggest a relationship between these two quantities. Indeed, the two conditions with the highest pupil dilation (6 and ±2 ST) are also the conditions with the highest entropy (Fig. 2*C*), while the condition with the least pupil dilation (0 ST) is also the condition in which entropy is lowest (responses are most predictable).

We confirmed this relationship at the individual level by regressing for each participant pupil size against entropy at each sample between 1 and 4 s after stimulus onset. Comparing the regression coefficients against zero using a cluster-based permutation test identified a significant cluster between 1.78 and 3.96 s (total t = 308.41, $P = 0.017$), with positive weights indicating an effect of response entropy on pupil size (Fig. 3 *A, C,* and *E*).

However, in our experimental design, the size of the log–frequency shift between the two tones covaried with both entropy and pupil size. Indeed, when we replicated the previous analysis with interval size instead of entropy, we also found a significant cluster between 1.8 and 4 s (total t = 327.04, $P = 0.009$; see Fig. 3 *B, D,* and *F*). To evaluate which of these two possible quantities (entropy versus interval size) best accounted for pupil dilation, we compared two different Bayesian hierarchical models, in which the mean pupil dilation in the condition cluster (2.06 to 4 s) was predicted from entropy or interval size. Both models significantly outperformed the null model with only an intercept per participant (entropy: $\Delta\text{DIC} = -29$, $P < 0.001$; interval size: $\Delta\text{DIC} = -9$, $P = 0.011$). Importantly, the entropy model also explained pupil size better than the interval size model ($\Delta\text{DIC} = -20$, $P < 0.001$).

In sum, in Experiment 1, we found that conditions associated with greater variability in perceptual decisions (higher entropy) also elicited greater pupil size, even though participants were completely unaware of the ambiguity of the stimulus in the case of 6 ST Shepard tones. Interval size may have been a possible confounding factor for the effect of entropy on pupil size, but model comparison provided some evidence favoring an explanation of the data based on entropy.

Experiment 2: Comparing Ambiguous and Nonambiguous Tone Shifts. ***Behavioral evidence for unaccessed ambiguity.*** In Experiment 2, we sought to confirm our main finding that pupil size would signal unaccessed ambiguity, while introducing three modifications

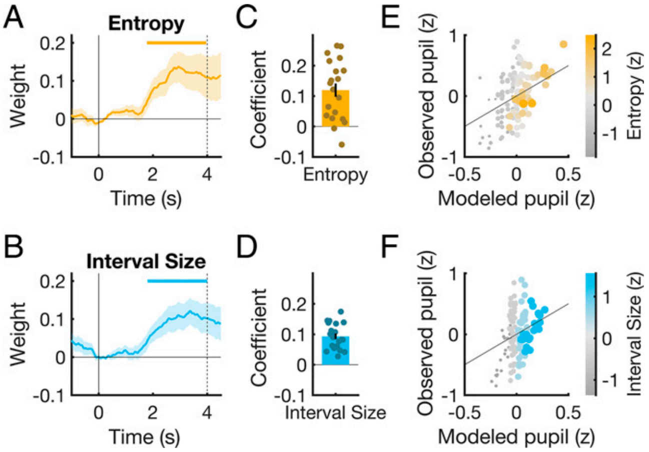

Fig. 3. Pupil dilation correlates with response entropy and interval size for ambiguous frequency shifts. (*A*) Time course of the entropy-related pupil response in Experiment 1, shown as beta weights from sample-by-sample regressions of response entropy on pupil size. The horizontal bar indicates $P < 0.05$ in a cluster-corrected permutation test against zero, and the colored region shows ±1 SEM ($n = 20$). (*B*) Time course of the interval size-related pupil response, as in *A* but replacing response entropy with interval size. (*Center*) Individual coefficients for response entropy (*C*) and interval size (*D*) in two separate Bayesian hierarchical models of mean pupil dilation from 2.06 to 4 s in Experiment 1. (*Right*) Comparison of predictions from the response entropy (*E*) and interval size (*F*) models and observed data. Each subject's individual mean has been subtracted from both model predictions and observed data in order to focus on within-subject differences. Each circle shows one condition for one listener, with the associated response entropy or interval size shown by both the color and the size of the circle.

in the design. First, awareness of ambiguity was now evaluated more explicitly, via confidence ratings on a scale from 1 to 4, following up versus down judgments. Second, we only tested small intervals (difference limens [DL]), medium (2.5 ST), and large (5.5 ST) intervals. Since the 0 ST and ±2 ST conditions were only introduced in Experiment 1 to confirm the valid use of the "both" and "neither" response options, we no longer need these conditions in Experiment 2 and could thus have a more homogeneous stimulus set in which there was always one and only one frequency shift to judge in all conditions. Third, to dissociate pupil size from interval size, we tested not only Shepard tones but also harmonic complex tones, which unlike Shepard tones should produce no ambiguity (and therefore minimal pupil dilation) in the large-interval condition (Fig. 4 *A* and *B*).

Behavioral results are shown in Fig. 4 *C* and *E*. As anticipated, the design of Experiment 2 dissociated response variability from interval size. Indeed, response variability decreased with interval size for harmonic tones [$F(1,24) = 13.14$, $P < 0.001$] but increased with interval size for Shepard tones [$F(1,24) = 23.37$, $P < 0.001$]. A two-way, repeated-measures ANOVA confirmed the tone type × interval size interaction on response variability [$F(2,48) = 47.46$, $P < 0.001$] and indicated a main effect of tone type [$F(1,24) = 17.73$, $P < 0.001$] but no main effect of interval size. Importantly, post hoc comparisons at each interval (Bonferroni corrected $\alpha = 0.017$) indicated greater response variability for Shepard than harmonic tones at 5.5 ST but not in the other conditions. This shows that we could manipulate response variability while keeping interval size constant at 5.5 ST when contrasting our two types of stimuli.

In line with Experiment 1 and with prior research (10), participants' unawareness of the ambiguous nature of large-interval Shepard tones was confirmed as well, with high confidence expressed for this condition (Fig. 4 *D* and *F*). For harmonic tones, confidence was also high for the large-interval condition, as expected since the pitch shift is subjectively large. The two-way ANOVA on confidence ratings confirmed the main effect of interval size [$F(2,48) = 13.58$, $P < 0.001$], with no main effect of tone type, but also indicated a tone type × interval size interaction [$F(2,48) = 7.98$, $P = 0.001$]. Post hoc pairwise comparisons between intervals for each tone type (Bonferroni-corrected $\alpha = 0.0083$) indicated significant differences between all intervals for harmonic tones, whereas for Shepard tones confidence was lower for the DL condition but similar between 5.5 ST and 2.5 ST conditions.

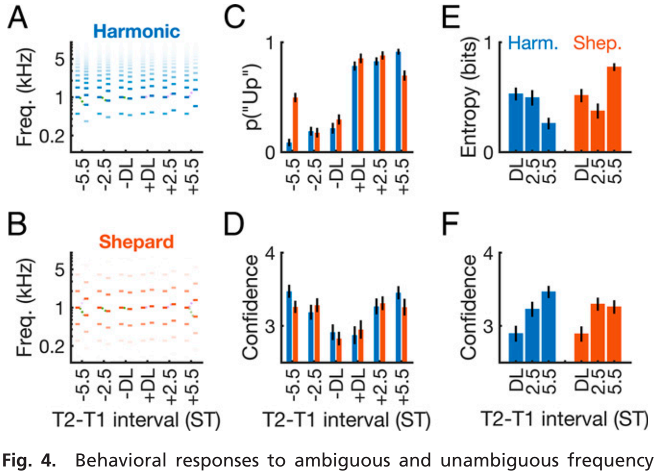

Fig. 4. Behavioral responses to ambiguous and unambiguous frequency shifts. (*A* and *B*) Schematic spectrograms of stimuli in Experiment 2, with dashed lines indicating shortest paths between components. The DL is the listener's individually measured frequency DL. Ambiguity emerges at 5.5 ST for Shepard tones but not for Harmonic complexes. (*C* and *D*) Probability of responding "up" and average confidence rated on a four-point scale for each condition in Experiment 2. Behavioral variability is greatest for 5.5 ST Shepard tone intervals and smallest for 5.5 ST harmonic complexes. Confidence is lowest at the DL for both tone types. Colored regions show ±1 SEM ($n = 25$). (*E* and *F*) Response entropy and confidence for each tone type and interval size, collapsing across interval sign. Error bars show ±1 SEM ($n = 25$).

Note that although confidence was blind to the ambiguity of large-interval Shepard tones, we could verify that it was not blind to fluctuations of performance in general. Indeed, a three-way ANOVA on confidence considering accuracy, interval size, and tone type found not only a main effect of interval size [$F(2,48) = 19.59$, $P < 0.001$] but also a main effect of accuracy, with higher confidence for correct responses than for errors [$F(1,24) = 48.86$, $P < 0.001$]. There was also a type × accuracy interaction [$F(1,24) = 5.20$, $P = 0.032$], reflecting a stronger variation of confidence with accuracy for harmonic complexes than for Shepard tones. No other significant main interactions or effects were observed in this ANOVA. In sum, this suggests that participants' confidence ratings truly carried information about the accuracy of their perceptual decisions above and beyond the variation of interval size and tone type.

Pupil dilation reflects both implicit ambiguity and explicit confidence. Despite the inability of participants to acknowledge the ambiguity in the stimulus for large-interval Shepard tones, pupil size was greater in this condition compared with other conditions (Fig. 5). Indeed, our cluster-based permutation approach revealed a significant cluster in which pupil size was sensitive to the tone type × interval size interaction (2.1 to 2.96 s, total F = 172.63, $P = 0.04$). There was also a cluster for the main effect of tone type (2.52 to 4 s, total F = 382.23, $P = 0.05$) but no cluster for interval size. Critically, pairwise comparisons within the interaction cluster (Bonferroni-corrected $\alpha = 0.017$) revealed that Shepard tones evoked greater pupil dilation than harmonic complexes at 5.5 ST but not at 2.5 ST or DL. Variations of pupil size across conditions thus mirrored the variations of behavioral variability reported in the previous section: Both pupil size and behavioral variability were maximal for ambiguous Shepard tones. The results of Experiment 2 thus replicate the finding of Experiment 1 that pupil size carries a signal about stimulus ambiguity even in the absence of subjective awareness about this ambiguity.

Measuring confidence in Experiment 2 allowed us to evaluate the suggested relationship between confidence and pupil size in our data. To do so, we conducted separate regressions of pupil size against the *z*-scored values of entropy, inverse confidence, and interval size (Fig. 6). The resulting regression weights were then compared against zero using cluster-based permutation tests. A significant cluster was identified for entropy (2.9 to 4 s, total t = 140.29, $P = 0.047$) but not for either of the other two predictors. Weights for inverse confidence were positive consistently with past studies (20); however, for confidence, no cluster could reach significance.

The goal of our last analysis was to select the best factors affecting pupil size in Experiment 2. To do so, we used a Bayesian model selection procedure, starting from the full model (entropy, confidence, interval size, tone type, and all interactions) and eliminating the weakest effects one by one until the Deviance Information Criterion (DIC) of the resulting model was at least +6 relative to the previous model, indicating a significant loss of information. The final model produced by this selection procedure contained the main effects of both entropy and (inverse) confidence, suggesting that both predictors contributed to the pupil response (Fig. 7). The final model also included interactions of entropy by size (with a stronger effect of entropy for large intervals) and of confidence by tone type (with a stronger effect of confidence for Shepard tones). This model was significantly more likely than either a null model containing only an intercept effect ($\Delta\text{DIC} = -28$, $P < 0.001$) or

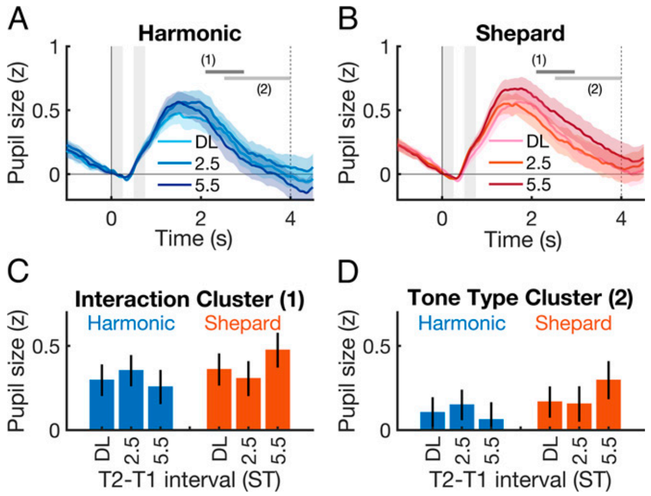

Fig. 5. Pupil dilation in response to ambiguous and unambiguous frequency shifts. (*A* and *B*) Time course of pupil dilation in Experiment 2, baseline-corrected, *z*-scored, and averaged across subjects in each condition. Gray regions indicate T1 and T2 presentation. The dashed vertical line indicates the onset of the response window. Gray horizontal bars indicate $P < 0.05$ in a cluster-corrected permutation test, in dark gray (1) for the interaction of interval size by tone type on pupil size, and in light gray (2) for the main effect of tone type on pupil size. Colored regions show ±1 SEM ($n = 25$). (*C* and *D*) Pupil size in each condition averaged within the interaction cluster (1) and the tone type cluster (2). Error bars show ±1 SEM ($n = 25$).

the full model with all possible effects and interactions (ΔDIC $= -15, P < 0.001$).

Discussion

In two experiments, we presented ambiguous auditory stimuli, namely large-interval Shepard tones, to naïve listeners. Behaviorally, the ambiguity of these stimuli was evidenced by the fact that they were equally likely to be perceived as sounds going up or down in pitch. However, in a given trial, participants did not perceive both directions but only one (Experiment 1), with a high level of confidence (Experiment 2). In other words, listeners were not aware of the intrinsic ambiguity in the stimulus that caused subsequent behavioral variability in their responses. In both experiments, however, pupil dilation was sensitive to this nonconsciously perceived ambiguity: pupil size increased more for ambiguous stimuli than for any other stimuli. Quantifying stimulus uncertainty as the entropy of the behavioral responses over the course of the experiment, we found that pupil dilation was correlated with stimulus uncertainty, independently of participants' subjective awareness of this uncertainty on any given trial. We now interpret this main result by considering various factors known to affect pupil dilation.

Pupil Dilation Does Not Only Reflect Perceived Qualitative Differences between Stimuli. A first potential confound that is important to rule out from the outset is that, in Experiment 2, Shepard tones and harmonic complex tones had a different timbre—they were audibly different to listeners. Moreover, Shepard tones sound less similar to natural sounds than harmonic complex sounds. If Shepard tones simply sounded "odd" to the listeners, this could have induced a greater pupil dilation for them (36). Critically though, this difference in timbre was present for all conditions, ambiguous or not. Thus, it cannot explain the key finding that pupil dilation was highest only for the ambiguous, large intervals made of Shepard tones and not, for instance, for the same but unambiguous, large intervals made of harmonic complex tones. In other words, such a confound could explain a main effect of tone type, but it could not explain the observed interaction between tone type and interval size. Therefore, pupil dilation was not only due to perceivable differences across stimulus categories but rather reflected stimulus ambiguity in all cases, with or without subjective awareness of this ambiguity.

Pupil Dilation Does Not Only Reflect Cognitive Effort. A second possible interpretation of the results could be sought in the classic findings relating pupil dilation to cognitive effort [(12, 37); see ref. 36 for a review in the auditory domain]. In the present case, it may be argued that larger pupil dilation for ambiguous sounds reflected a greater listening effort or higher processing load for such stimuli. A similar idea has indeed been explored for the lexical ambiguity between two different meanings of a word, which may require more effort (38) and indeed elicit greater pupil dilation (39). A key difference with our study, however, is that the two competing interpretations are well known to participants in the case of lexical ambiguity, whereas they are not in the case of ambiguous Shepard tones. More generally, we would argue that an interpretation in terms of effort alone cannot fully account for our data for two reasons. Firstly, had ambiguous Shepard tones imposed a higher cognitive load or a greater listening effort, one would expect also lower confidence judgments for these stimuli. This was not the case in our data nor was it the case in a previous study in which response times were also collected but showed no

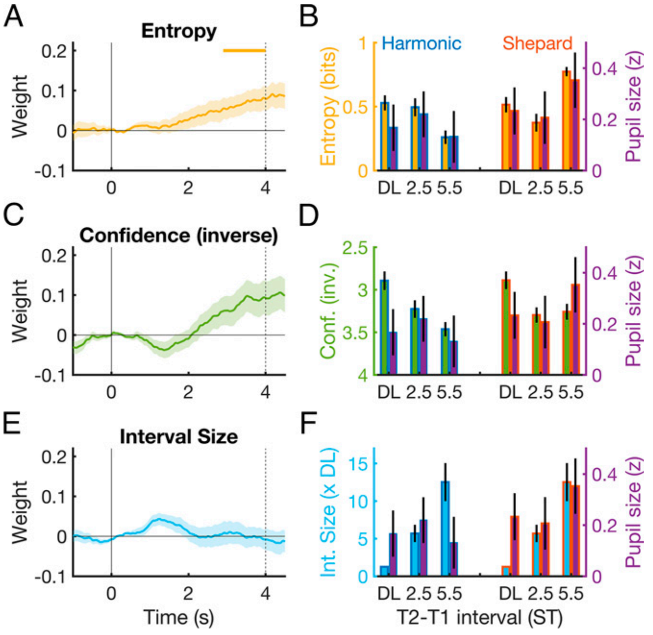

Fig. 6. Pupil dilation correlates with response entropy. (*A*) Time course of the entropy-related pupil response in Experiment 2, shown as beta weights from sample-by-sample regressions of response entropy on pupil size. The horizontal bar indicates $P < 0.05$ in a cluster-corrected permutation test against zero, and the colored region shows ±1 SEM ($n = 25$). (*B*) Comparison across conditions of response entropy and mean pupil dilation in the overall cluster previously identified in Experiment 1 (2.06 to 4 s). Error bars show ±1 SEM. (*C*) Time course of the confidence-related pupil response, as in *A* but replacing response entropy with confidence ratings, inverted to reflect the expected direction of the effect. No significant cluster was identified. (*D*) Comparison across conditions of inverted confidence and pupil dilation, as in *B*. (*E*) Time course of the interval size–related pupil response, as in *A* but replacing response entropy with interval size. No significant cluster was identified. (*F*) Comparison across conditions of interval size and pupil dilation, as in *B*.

https://doi.org/10.1073/pnas.2107997118

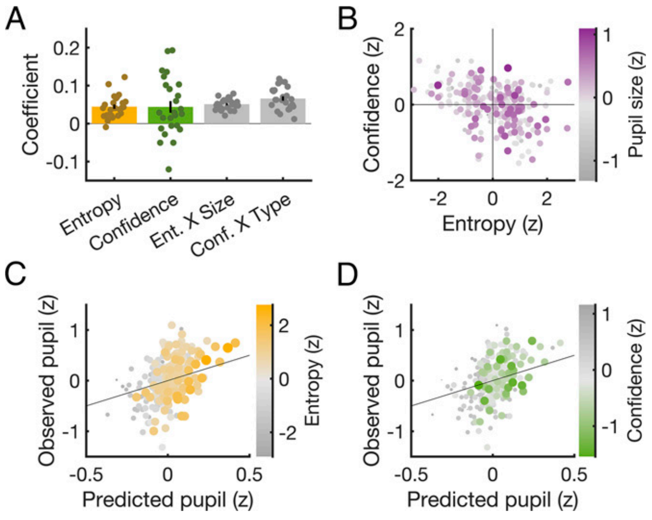

Fig. 7. Modeling the effects of response entropy and confidence on pupil dilation. (*A*) Individual coefficients for each of the four effects in a Bayesian hierarchical model of mean pupil dilation from 2.06 to 4 s in Experiment 2. Error bars show ±1 SEM (*n* = 25). (*B*) Comparison of response entropy and confidence in Experiment 2. Each circle shows one condition for one listener, with the associated mean pupil dilation shown by both the color and size of the circle. Response entropy and confidence are negatively correlated (r = −0.37), but examples of both high-entropy/high-confidence and low-entropy/low-confidence combinations are numerous. (*C*) Comparison of predictions from the Bayesian hierarchical model and observed data. Each subject's individual mean has been subtracted from both model predictions and observed data, in order to focus on within-subject differences. Each circle shows one condition for one listener, with the associated response entropy shown by both the color and the size of the circle. (*D*) The same as in *C* but with color and size of circles showing confidence rather than response entropy.

evidence of a greater effort for ambiguous Shepard tones (10). Secondly, in our second experiment, intervals presented at the listener's DL arguably require more listening effort in terms of discrimination, but they were not those for which pupil dilation was the largest. Importantly, this does not imply that effort plays no role in our data, only that it does not constitute a simple and complete explanation.

Pupil Dilation Does Not Only Reflect Metacognitive Confidence. A third framework to interpret our findings is related to the notion of confidence, which has been shown to be reflected in pupil size (e.g., refs. 19 and 20). Several points can be made in regard to this.

First of all, it is important to note that entropy and confidence index distinct facets of the notion of uncertainty. Confidence, as we operationalized it, is a measure of conscious uncertainty, namely a form of uncertainty that is subjectively felt and can be reported by participants. Entropy, on the other hand, is a measure of objective, behavioral uncertainty, which may or may not be reflected to consciousness. In typical perceptual tasks with near-threshold stimuli, confidence and entropy are negatively correlated. However, when ambiguous stimuli are introduced, these two factors can be dissociated, since the ambiguous condition may, paradoxically, be associated with both high confidence and high entropy. In Experiment 2, we found in particular that although entropy and confidence were only weakly correlated, both factors contributed to pupil dilation.

Secondly, the high confidence observed in Experiment 2 comports with the decoupling of explicit reports of confidence from the behavioral variability observed in a previous study of ambiguous, Shepard-tone, 6 ST intervals (10): While trained musicians and nonmusicians showed the same behavioral variability in their responses, nonmusicians showed increased confidence whereas musicians show decreased confidence in their responses to the ambiguous stimulus. In the present study, the majority of participants included listeners with no or little musical experience, who showed high confidence in the ambiguous condition. Thus, the stronger association between entropy and pupil response reveals that already in naïve listeners we find a neural signature of ambiguity processing. How this neural signal of ambiguity changes with expertise and possibly reaches awareness constitutes a question for future research.

Thirdly, although response times are sometimes considered as an implicit measure of confidence, they may not provide a good account of pupil dilation in the ambiguous condition in our study. In our previous study, the ambiguous condition was not associated with longer response times in untrained listeners (10). Besides, for nonambiguous stimuli, pupil dilation correlates with evidence strength even when controlling for response time (21). Our paradigm did not allow for an informative measurement of response time, since listeners had to wait several seconds before responding, likely well after a perceptual decision has been made. Future studies could modify the paradigm to allow quick responses and compare the two measures for sensitivity to this ambiguity.

Pupil Dilation and the Implicit Processing of Stimulus Uncertainty. In light of the alternative interpretations discussed so far and their limitations, we argue that a specific and parsimonious interpretation of our results is that, even when listeners were not conscious of it, the uncertainty associated with an ambiguous stimulus was reflected in their pupillary response. We may say that the eye knows something that the I doesn't: In implicitly ambiguous conditions, participants lacked metacognitive access of their variable behavior, but the pupil tracked it.

A number of previous studies have proposed that pupil dilation is sensitive to uncertainty, as outlined in the introduction (15–21, 28), with various definitions of uncertainty, from reward structure to stimulus identity to internal parameters of probabilistic models. Here, we operationalized stimulus uncertainty as the variability of behavioral responses to repeated presentations of that stimulus. Importantly, we argue that, although measured across trials in our experimental setting, uncertainty is in fact inherent to the stimulus itself. Supporting this view, we show that pupil dilation varies on a trial-by-trial basis as a function of this measure. More generally, even for single presentations, the sensory representation of a stimulus will necessarily contain some degree of uncertainty (3). Consistently, with recent accounts of entropy that emphasize the probabilistic character of neural representations (40), it seems likely that the brain is simply maintaining probabilistic representations of the environment that collapse, at some later level, into one interpretation.

Our findings extend previous studies concerned with pupillary response to bistable stimuli (31, 32), as bistable stimuli exhibit high stimulus uncertainty under our definition. In these studies, pupil dilation was observed around the time of the perceptual switches reported by observers. Einhäuser et al. (31) proposed that, consistent with current models of bistability, ambiguity resolution involved a competition between alternative interpretations of the stimulus along the perceptual pathways (see e.g., ref. 41). They then argued that the neuromodulatory activity reflected in pupil dilation was a gain signal consolidating the interpretation emerging from the competition, thus favoring the emergence of a stable percept. However, Hupé et al. (32) pointed out that instead of revealing mechanisms involved in resolving ambiguity, pupil dilation may rather reflect the consequence of this resolution, such as awareness of a perceptual change or subsequent motor responses. Since our participants were neither told of the potential ambiguity of the stimulus, nor asked to wait for a switch in their percept (see ref. 42), and since they were unaware of the stimulus uncertainty, our paradigm does not suffer from the potential confounds raised by Hupé et al. (32). In other words, rather than reflecting the consequence of ambiguity resolution, our results may indeed help characterize its neural underpinnings.

This allows us to reconsider the proposed role of neuromodulation in the resolution of perceptual ambiguity and uncertainty more generally. For instance, neuromodulatory signals reflected in pupil dilation could facilitate the emergence of a stable interpretation of the stimulus in rivalrous conditions through gain modulation (31). Other proposals have related neuromodulators to network resets (43), the regulation of exploration versus exploitation (25, 44, 45), or the coordination of neural activity across brain regions (46). All of these functional roles may be useful to ambiguity resolution. Our data do not specifically arbitrate between these proposals but show that neuromodulation may operate independently of the subjective awareness of the uncertainty. It remains an open question how stimulus ambiguity becomes accessible to consciousness and why it sometimes remains altogether hidden from awareness.

Materials and Methods

Experiment 1.

Listeners. A total of 20 adult listeners, 13 female and 7 male and 16 right handed and 4 left handed, were recruited to participate in the study. The number of participants was determined before data collection using a power analysis (available at https://osf.io/skdfp). In audiometric screening, 19 listeners had thresholds at or below 20 dB hearing level (HL) for octave frequencies 250 to 8,000 Hz, and one listener had thresholds at or below 35 dB HL for these frequencies. Listeners ranged in age from 20 to 49 y old (M = 27.45, SD = 6.92) and reported an average of 2.95 y of musical training, with 12 out of 20 listeners reporting no musical training at all. When asked to rate their level of musical ability on a scale from 1 to 10, listeners rated themselves at a mean of 2.43 (SD = 2.01). All listeners gave their written informed consent to participate in the experiment, in accordance with the Declaration of Helsinki and local ethics procedures. The experiment was approved by a local ethics committee (Conseil en éthique pour les recherches en santé [CERES], institutional review board [IRB] approval 20154000001072).

Stimuli and apparatus. The auditory stimulus on each trial was composed of two consecutive, complex tones, T1 and T2, each with a duration of 250 ms, including 10-ms on- and off-ramps. Tones were separated by a silent interval of 250 ms, such that T2 onset occurred 500 ms after T1 onset. All tones were embedded in threshold-equalizing noise (47) at 40 dB sound pressure level (SPL) within the equivalent rectangular bandwidth (ERB) centered around 1 kHz, which was 10 dB below the maximum level for a single-frequency component of a tone (50 dB SPL). The delay between noise onset and T1 onset randomly varied from 3 to 5 s on each trial, and the offset of the noise always occurred 4 s after the onset of the first tone.

In most experimental conditions, both complex tones were Shepard tones (8), consisting of a base frequency multiplied by every integer power of two, such that components are spaced at octaves and extend above and below the base frequency. A constant spectral envelope was applied to the components of each tone regardless of base frequency, taking the form of a Gaussian distribution with a centroid of 1 kHz and an SD of 1 octave. Tones were scaled such that the maximum presentation level of a component within this spectral envelope was 50 dB SPL.

See Fig. 1*B* for a schematic depiction of the different experimental conditions in Experiment 1. The distance separating the base frequencies of the two tones varied depending on the experimental condition: The base frequency of T2 was either 0, 2, 4, 6, 8, or 10 ST higher than the base frequency of T1. The base frequencies of the two tones were centered around one of four center frequencies: 440 Hz (A4), 523 Hz (C5), 622 Hz (D#5), or 740 Hz (F#5). These four center frequencies were presented in equal number to each participant in each experimental condition in a randomized order.

One experimental condition, termed ±2 ST, differed qualitatively from all the other conditions in terms of stimulus construction. The first tone in this condition was a Shepard tone, as in all other conditions. The frequency components of the second tone were then chosen according to the following rule: Components an odd number of octaves away from the base frequency were shifted down by 2 ST, while components an even number of octaves away from the base frequency were shifted up by 2 ST. In this way, the amount of evidence for an upward frequency shift is equal to the amount of evidence for a downward frequency shift, as in the 6 ST condition.

All sounds were generated within MATLAB (The Mathworks) using a 24-bit RME Digiface USB soundcard (Audio AG) and were presented diotically through DT770 headphones (Beyerdynamic) at a sampling rate of 44.1 kHz. Pupillometric data were recorded from both eyes simultaneously using a Pupil Core eye tracker (Pupil Labs) at a sampling rate of 200 Hz.

Procedure. Listeners first completed a brief training and orientation procedure in which the behavioral task was explained and demonstrated to them without pupillometric recording. Listeners were instructed to characterize the direction of the change between the two tones presented on each trial using one of four response options: "up," "down," "both," or "neither," coded as the up, down, right, and left arrow keys, respectively. Listeners were asked to use the "up" or "down" responses if they heard only a single direction of change, the "both" response if they heard evidence for two directions of change on a single trial, and the "neither" response if they thought there was no change between the two tones. Stimuli during the training procedure were as in the full test, except that tones were harmonic complexes containing all integer multiples of the fundamental frequency (F0), which eliminates the ambiguity inherent in Shepard-tone stimuli. In the training ±2 ST condition, T2 consisted of two simultaneous harmonic complexes, with F0s at +2 ST and −2 ST relative to the F0 of T1. In a randomized order, listeners completed four trials each of four training conditions: −2 ST, 0 ST, +2 ST, and ±2 ST for a total of 16 trials.

After the training procedure, listeners completed the full test while pupil diameter was continuously recorded from both eyes. During the full test, listeners were instructed to fixate on a cross in the center of the screen and to wait until 4 s after T1 onset before responding on each trial. The 4-s mark was indicated by the offset of masking noise, as well as a change in the color of the fixation cross. Listeners completed a total of 120 trials in the full test, with experimental conditions presented in a randomized order. Each listener completed a total of 24 trials each for the 0 ST, 6 ST, and ±2 ST conditions and 12 trials each for the 2, 4, 8, and 10 ST conditions. The 2 and 10 ST conditions, as well as the 4 and 8 ST conditions, produce perceptually equivalent stimuli with dominant shifts in opposite directions. When these pairs of conditions are combined, there are 24 trials in each condition.

Quantifying variability in behavioral responses with entropy. We used entropy (48) as an objective measure of behavioral variability in each condition:

$$H = -\sum_{r=1}^{4} p_r \log_2 p_r$$

where H is the entropy or empirical uncertainty in the behavioral response, p_r is the probability of one response type [e.g., p(up)] for one subject in one condition, and r varies from 1 to 4 for the four response types. A condition in which one response was chosen 100% of the time would have 0 bits of entropy (the minimum), while a condition in which each of the four responses was chosen 25% of the time would have 2 bits of entropy (the maximum). This classical measure of entropy can be interpreted as variability, uncertainty, surprisal, or information (40). In a condition with 0 entropy, the responses are invariable, certain, unsurprising, and contain no information; whereas in a condition with 2 bits of entropy, the responses are variable, each response is uncertain and surprising, and conveys information.

Preprocessing of pupil data. Pupil data were preprocessed broadly according to guidelines described by Kret and Sjak-Shie (49). First of all, samples for which Pupil Labs' reported confidence at acquisition was less than 90% were rejected. Then, a step-by-step process marked artifacts for rejection: First, the 5% of data with most extreme (large or small) absolute pupil diameter was rejected. Next, the 20% of data with the largest recorded eye gaze eccentricity was rejected. Next, the 20% of data with the highest instantaneous dilation speed was rejected. Then, "islands" of continuous data lasting less than 100 ms, and separated from other data by at least 75 ms, were rejected. Then, for each gap in the data between 75 and 2,000 ms long, the 50 ms before and after the gap were also rejected. Then, missing data were interpolated, and the trendline was estimated using a fourth-order Butterworth low-pass filter with a cutoff of 2 Hz, and the 20% of data that deviated most from this trendline was rejected. Island rejection, gap padding, and trendline-based rejection were repeated 20 times, and after the 20th iteration, gaps larger than 250 ms were labeled as missing data, while smaller gaps were interpolated. Data were then low-pass filtered (2 Hz, fourth-order Butterworth) and separately z-scored in each eye. Finally, data from both eyes were averaged together, and the low-pass filter and z-scoring were reapplied a final time.

Data were initially recorded at 200 Hz, which is the maximum available for the Pupil Labs device and is sufficient to measure changes in pupil size, which generally occur at much lower frequencies. Sampling rates of 200 Hz or lower have previously been used to measure pupil size (16, 17, 32), and for pupil dilation data, these slower sampling rates have been directly compared and found to be equivalent with the 1-kHz rate commonly used to measure eye movements (50, 51). Because of the slow time scale of pupil effects, the low-pass filter applied at the end of preprocessing used a standard cutoff of 2 Hz, removing energy at higher frequencies. In order to reduce the computational load for later cluster-based time series analyses, and to facilitate plotting, the data were downsampled to 50 Hz after all preprocessing was complete.

The preprocessed stream of pupil data were divided into epochs relative to the onset of T1 on each trial. A critical period was defined between 100 ms prior to T1 onset and 4,000 ms after T1 onset. Trials with too much missing data were then eliminated using the following procedure: First, any trial with more than 15% missing data in the critical period was eliminated. However, if this resulted in more than 8 out of 24 trials being eliminated in any condition, the criterion was relaxed for that condition to <50% missing data in the critical period. The relaxed criterion was used in at least one condition for 15 subjects in experiment 1 and for eight subjects in experiment 2. For each condition and subject, a mean of 16.9 trials were retained in experiment 1 (with at least 14 trials retained in 90% of cases) and 19.4 trials in experiment 2 (with at least 16 trials retained in 90% of cases).

The pupil signal on each trial epoch was baseline corrected. Specifically, the mean over 600-ms period starting 100 ms before T1 onset and ending at T2 onset was subtracted from the entire epoch. Pupil signals in each condition were then averaged together to compute a pupil trace for each subject in each condition. These individual pupil traces are shown in *SI Appendix*, Fig. S1 for Experiment 1.

Identifying significant effects on pupil dilation. The time course of the main effect of condition on pupil dilation was measured by computing sample-by-sample, one-way, repeated-measures ANOVAs of condition on pupil dilation and recording the F statistic at each sample between 1 and 4 s. A period of significance in this time series of F statistics was identified using cluster-based permutation statistics (52). The F statistics were summed for the largest consecutive series (cluster) of values $F > 3$, and this F sum was compared with the F sum of the largest $F > 3$ cluster on 10,000 iterations of the same process, after random permutation of conditions for each subject.

The time course of the effects of response entropy and interval size on pupil dilation were measured by computing sample-by-sample, independent linear regressions. Beta weights from the regressions were compared with zero using *t* tests, and the t statistic at each sample between 1 and 4 s was recorded. Periods of significance in these time series of t statistics were identified by comparing the largest cluster of values $t > 2$ with the equivalent largest cluster on 10,000 iterations of the same process, with random permutations of sign for each subject.

Modeling effects of response entropy and interval size on pupil dilation. A significant cluster for the effect of condition on pupil dilation in Experiment 1 was identified at 2.06 to 4 s (see *Results*, *Experiment 1*: *Response Variability and Pupil Dilation for Ambiguous Tone Shifts*). We modeled the variance in the mean pupil dilation over this period by fitting two different versions of a Bayesian hierarchical model, implemented in R using the rjags package (53). This class of model separately estimates each effect for each subject and also estimates the hyperparameters of the mean and variance of the distributions for each effect. The parameters of the model were estimated using the Markov chain Monte Carlo (MCMC) method in rjags, with three chains initialized with 20,000 iterations each, and posterior distributions estimated over 10,000 iterations. One version of the model contained a single fixed effect of entropy:

$$P = \beta_0 + \beta_H \times H + \varepsilon$$

where P is pupil size, H is entropy, β_0 is the intercept, β_H is the effect of entropy, and ε is an error term. The other version of the model replaced the effect of entropy with an effect of interval size:

$$P = \beta_0 + \beta_S \times S + \varepsilon$$

We compared the two models using the DIC, a generalization of the Akaike Information Criterion (AIC) applicable to MCMC Bayesian hierarchical models (54). Like the AIC, the DIC is a measure of model performance that penalizes complexity and can be used to compare models (55). The relative likelihood L of one model compared with another model is given by:

$$L = e^{\frac{DIC_2 - DIC_1}{2}}$$

The two versions of the model using entropy and interval size were compared with each other on the basis of DIC-derived likelihood, and each version of the model was also compared against a null model containing only the intercept and error terms.

Experiment 2.

Listeners. A total of 25 adult listeners, 18 female and 7 male and 24 right handed and 1 left handed, were recruited to participate in the study. The number of participants was determined before data collection using a power analysis (available at https://osf.io/nx8tp/); however, because of experimenter error, the number of subjects recruited was two fewer than recommended by this power analysis. Recruitment was stopped at 25 subjects before any data analysis was performed, and the resulting reduction in statistical power was negligible (from 95 to 93%). In audiometric screening, 23 listeners had thresholds at or below 20 dB HL for octave frequencies 250 to 8,000 Hz, and two listeners had thresholds at or below 35 dB HL for these frequencies. Listeners ranged in age from 19 to 47 y old (M = 28.48, SD = 7.04) and reported an average of 3.45 y of musical training, with 12 out of 25 listeners reporting no musical training at all. When asked to rate their level of musical ability on a scale from 1 to 10, listeners rated themselves at a mean of 3.00 (SD = 2.10). All listeners gave their written informed consent to participate in the experiment, in accordance with the Declaration of Helsinki and local ethics procedures. The experiment was approved by a local ethics committee (CERES, IRB approval 20154000001072).

Stimuli and apparatus. As in Experiment 1, the auditory stimulus on each trial was composed of two consecutive complex tones embedded in noise. Most details of stimulus generation and presentation were identical to Experiment 1, differing only in the size of the frequency shift between tones and the introduction of harmonic complex tones.

On half of all trials, both tones were Shepard tones as in Experiment 1, with frequency components spaced at octaves extending above and below the base frequency (FB). On the other half of trials, both tones were harmonic complex tones, consisting of a fundamental frequency (F0) and all positive integer multiples of F0. The same Gaussian spectral envelope was applied to both tone types, with a centroid of 1 kHz and an SD of 1 octave, with the maximum level of a single component set to 50 dB SPL.

See Fig. 4 *A* and *B* for a schematic depiction of the different experimental conditions in Experiment 2. Each trial fell into one of three interval size categories: DL, 2.5 ST, or 5.5 ST and consisted of either Shepard tones or Harmonic complex tones. The FB (Shepard tones) or F0 (Harmonic complexes) at the geometric center of the two tones was randomly sampled on each trial from a uniform distribution of logarithmic frequency on the octave from 260 to 520 Hz (C4 to C5). The FB or F0 of each tone was then determined based on the interval size for that trial. Within each tone type and interval size category, 12 specific interval sizes were defined, evenly spaced from 5 to 6 ST for the 5.5 ST category, from 2 to 3 ST for the 2.5 ST category, and from 0.5 to 2 times the listener's individually measured threshold in the Threshold category. The variety of specific interval sizes was designed to prevent listeners from recognizing and remembering the musical quality of specific intervals from trial to trial.

Measurement of individual thresholds. Before participation in the main experiment, listeners' F0 DL were measured using a one-up, two-down adaptive tracking procedure, using only harmonic complex tones. Each listener completed three runs of an adaptive staircase with a starting interval size of 5.95% (1 ST). On each reversal, this interval size changed by a factor of 1.58, 1.26, 1.19, and finally 1.10, with the run completing after the sixth reversal at the final step sizes. The threshold on each run was estimated as the average of the last four reversals, and the three runs were averaged together to get a final threshold estimate. This estimate was used to define the stimuli in the Threshold condition of the main experiment for each individual listener. During threshold measurement, participants received feedback after each trial, showing whether their response had been correct or incorrect. This feedback was absent during the main experiment. Pupil size was not measured during threshold measurement.

Main experiment procedure. In the main experiment, on each trial, listeners heard two consecutive tones, waited until 4 s after tone 1 onset, and then made two separate behavioral responses. First, they were asked to say whether the direction of the change was up or down. Second, they rated their degree of confidence that the first response was correct, on a scale from 1 ("not all confident") to 4 ("extremely confident"). Pupil size was continuously measured during the main experiment.

Listeners completed a total of 144 trials in the full test, with experimental conditions presented in a randomized order. Each combination of tone type (Shepard or harmonic), direction (up or down), interval size category (DL, 2.5 ST, or 5.5 ST), and specific interval size (12 levels per category) was presented once. In this way, each listener completed 24 trials within each tone type and interval size category (12 specific interval sizes in each direction).

Identifying significant effects on pupil dilation. Preprocessing of pupillometric data was identical to Experiment 1. For purposes of condition-by-condition rejection of trials with too much missing data, trials were grouped only by tone type and interval size category, leaving 24 trials per condition. Individual pupil traces for each condition and subject in Experiment 2 are shown in *SI Appendix*, Figs. S2 and S3.

Cluster-based permutation tests were computed on three different kinds of time series, with all parameters identical to cluster identification in Experiment 1. Clusters were identified for the F statistics for each effect in a two-way, repeated-measures ANOVA on pupil dilation with effects of interval size (DL, 2.5 ST, or 5.5 ST) and tone type (Shepard and harmonic). Clusters were also identified for F statistics for separate, two-way ANOVAs on Shepard tones and harmonic complexes, with effects of interval size and mean–split confidence (high or low). Finally, clusters were identified against zero for beta weights of independent linear regressions for each of three predictors: entropy, confidence (inverted), and interval size.

Modeling contributions to pupil dilation. We considered the condition cluster previously identified in Experiment 1 (2.06 to 4 s) as our main analysis window for pupil size in Experiment 2. We modeled the variance in mean pupil dilation in this window using a Bayesian hierarchical model considering four independent variables: entropy (per condition), confidence (per trial), interval size (per trial, in terms of each listener's DL), and tone type (Shepard or harmonic). Each of these four variables was *z*-scored for each subject so that the units of modeled coefficients are equivalent across factors. In order to determine the appropriate model, we used a version of the model selection procedure described by Jaeger (56), wherein one factor at a time is removed from a large initial model, until the point in which removing the factor significantly degrades the model. In order to compare models in this procedure, we again used DIC, a measure of model fit that avoids overfitting by penalizing complexity. The reduced version of the model was chosen at each step unless the DIC of the reduced model was at least +6 relative to the previous version (i.e., unless the reduced model's likelihood was $P < 0.05$ relative to the previous model's likelihood). The factor chosen to be potentially removed at each step was the estimated population mean closest to zero. The initial model contained 15 total effects, comprising all possible effects: the four main effects, six two-way interactions, four three-way interactions, and the four-way interaction.

Data Availability. All behavioral and pupillometric data from both experiments in this article have been deposited in Open Science Framework (https://osf.io/634ua/?view_only=25755ba7a1d148-81b927bd6f792ec668). All of these data are analyzed and included in this article.

ACKNOWLEDGMENTS. We are grateful to the "Ambiguity - Production and Perception" group at the University of Tübingen for helpful feedback on this project. This work was funded by the grants "AMBISENSE" (Grant No. ANR-19-CE28-0019-01) and "VICONTE" (Grant No. ANR-18-CE28-0015) from the French Agence Nationale de Recherche. We also acknowledge Grant No. ANR-17-EURE-0017 FrontCog for research conducted at the Department of Cognitive Studies of the École Normale Supérieure, PSL University.

1. M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
2. Y. Weiss, E. P. Simoncelli, E. H. Adelson, Motion illusions as optimal percepts. *Nat. Neurosci.* **5**, 598–604 (2002).
3. D. Kersten, P. Mamassian, A. Yuille, Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
4. D. Rahnev, R. N. Denison, Suboptimality in perceptual decision making. *Behav. Brain Sci.* **41**, e223 (2018).
5. C. S. Pierce, J. Jastrow, On small differences of sensation. *Mem. Natl. Acad. Sci.* **3**, 73–83 (1884).
6. D. H. Brainard, A. C. Hurlbert, Colour vision: Understanding #TheDress. *Curr. Biol.* **25**, R551–R554 (2015).
7. D. Pressnitzer, J. Graves, C. Chambers, V. de Gardelle, P. Egré, Auditory perception: Laurel and Yanny together at last. *Curr. Biol.* **28**, R739–R741 (2018).
8. R. N. Shepard, Circularity in judgments of relative pitch. *J. Acoust. Soc. Am.* **36**, 2346–2353 (1964).
9. D. Deutsch, A musical paradox. *Music Percept.* **3**, 275–280 (1986).
10. C. Pelofi, V. de Gardelle, P. Egré, D. Pressnitzer, Interindividual variability in auditory scene analysis revealed by confidence judgements. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160107 (2017).
11. E. H. Hess, J. M. Polt, Pupil size in relation to mental activity during simple problem-solving. *Science* **143**, 1190–1192 (1964).
12. D. Kahneman, J. Beatty, Pupil diameter and load on memory. *Science* **154**, 1583–1585 (1966).
13. S. D. Goldinger, M. H. Papesh, Pupil dilation reflects the creation and retrieval of memories. *Curr. Dir. Psychol. Sci.* **21**, 90–95 (2012).
14. P. van der Wel, H. van Steenbergen, Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychon. Bull. Rev.* **25**, 2005–2015 (2018).
15. J. Qiyuan, F. Richer, B. L. Wagoner, J. Beatty, The pupil and stimulus probability. *Psychophysiology* **22**, 530–534 (1985).
16. M. R. Nassar *et al.*, Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat. Neurosci.* **15**, 1040–1046 (2012).
17. M. Quirins *et al.*, Conscious processing of auditory regularities induces a pupil dilation. *Sci. Rep.* **8**, 14819 (2018).
18. T. H. Muller, R. B. Mars, T. E. Behrens, J. X. O'Reilly, Control of entropy in neural models of environmental state. *eLife* **8**, e39404 (2019).
19. F. Meyniel, Brain dynamics for confidence-weighted learning. *PLOS Comput. Biol.* **16**, e1007935 (2020).
20. K. M. Lempert, Y. L. Chen, S. M. Fleming, Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLoS One* **10**, e0126588 (2015).
21. A. E. Urai, A. Braun, T. H. Donner, Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat. Commun.* **8**, 14637 (2017).
22. P. R. Murphy, R. G. O'Connell, M. O'Sullivan, I. H. Robertson, J. H. Balsters, Pupil diameter covaries with BOLD activity in human locus coeruleus. *Hum. Brain Mapp.* **35**, 4140–4154 (2014).
23. S. Joshi, Y. Li, R. M. Kalwani, J. I. Gold, Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron* **89**, 221–234 (2016).
24. J. Reimer *et al.*, Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nat. Commun.* **7**, 13289 (2016).
25. G. Aston-Jones, J. D. Cohen, An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
26. S. J. Sara, S. Bouret, Orienting and reorienting: The locus coeruleus mediates cognition through arousal. *Neuron* **76**, 130–141 (2012).
27. R. S. Larsen, J. Waters, Neuromodulatory correlates of pupil dilation. *Front. Neural Circuits* **12**, 21 (2018).
28. D. Kahnemann, J. Beatty, Pupillary responses in a pitch-discrimination task. *Percept. Psychophys.* **2**, 101–105 (1967).
29. J.-L. Schwartz, N. Grimault, J.-M. Hupé, B. C. J. Moore, D. Pressnitzer, Multistability in perception: Binding sensory modalities, an overview. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 896–905 (2012).
30. D. A. Leopold, N. K. Logothetis, Multistable phenomena: Changing views in perception. *Trends Cogn. Sci.* **3**, 254–264 (1999).
31. W. Einhäuser, J. Stout, C. Koch, O. Carter, Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1704–1709 (2008).
32. J.-M. Hupé, C. Lamirel, J. Lorenceau, Pupil dynamics during bistable motion perception. *J. Vis.* **9**, 10 (2009).
33. D. C. Knill, A. Pouget, The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
34. H. von Helmholtz, "The facts in perception" in *Epistemological Writings*, S. Hertz, Ed. (*Boston Studies in the Philosophy of Science Series*, Springer Netherlands, 1878), **37**, pp. 115–185.
35. J. D. Smith, W. E. Shields, D. A. Washburn, The comparative psychology of uncertainty monitering and metacognition. *Behav. Brain Sci.* **26**, 317–339 (2003).
36. A. A. Zekveld, S. E. Kramer, J. M. Festen, Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear Hear.* **32**, 498–510 (2011).
37. D. Kahneman, *Attention and Effort* (Prentice-Hall, 1973).
38. I. S. Johnsrude, J. M. Rodd, "Factors that increase processing demands when listening to speech" in *Neurobiology of Language*, G. Hickok, S. L. Small Eds. (Elsevier, 2016), pp. 491–502.
39. M. Kadem, B. Herrmann, J. M. Rodd, I. S. Johnsrude, Pupil dilation is sensitive to semantic ambiguity and acoustic degradation. *Trends Hear.* **24**, 2331216520964068 (2020).
40. M. Sprevak, Two kinds of information processing in cognition. *Rev. Phil. Psychol.* **11**, 591–611 (2020).
41. F. Tong, M. Meng, R. Blake, Neural bases of binocular rivalry. *Trends Cogn. Sci.* **10**, 502–511 (2006).
42. I. Rock, S. Hall, J. Davis, Why do ambiguous figures reverse? *Acta Psychol. (Amst.)* **87**, 33–59 (1994).
43. S. Bouret, S. J. Sara, Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends Neurosci.* **28**, 574–582 (2005).
44. M. Usher, J. D. Cohen, D. Servan-Schreiber, J. Rajkowski, G. Aston-Jones, The role of locus coeruleus in the regulation of cognitive performance. *Science* **283**, 549–554 (1999).
45. M. Jepma, S. Nieuwenhuis, Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *J. Cogn. Neurosci.* **23**, 1587–1596 (2011).
46. J. M. Shine, Neuromodulatory influences on integration and segregation in the brain. *Trends Cogn. Sci.* **23**, 572–583 (2019).

47. B. C. J. Moore, M. Huss, D. A. Vickers, B. R. Glasberg, J. I. Alcántara, A test for the diagnosis of dead regions in the cochlea. *Br. J. Audiol.* **34**, 205–224 (2000).

48. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

49. M. E. Kret, E. E. Sjak-Shie, Preprocessing pupil size data: Guidelines and code. *Behav. Res. Methods* **51**, 1336–1342 (2019).

50. M. Turi, D. C. Burr, P. Binda, Pupillometry reveals perceptual differences that are tightly linked to autistic traits in typical adults. *eLife* **7**, e32399 (2018).

51. B. V. Ehinger, K. Groß, I. Ibs, P. König, A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *PeerJ* **7**, e7086 (2019).

52. R. C. Blair, W. Karniski, An alternative method for significance testing of waveform difference potentials. *Psychophysiology* **30**, 518–524 (1993).

53. M. Plummer, rjags: Bayesian graphical models using MCMC. Version 4-12. https://mcmc-jags.sourceforge.io/ (Accessed 26 February 2020).

54. D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. Van Der Linde, Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 583–639 (2002).

55. E.-J. Wagenmakers, S. Farrell, AIC model selection using Akaike weights. *Psychon. Bull. Rev.* **11**, 192–196 (2004).

56. T. F. Jaeger, Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* **59**, 434–446 (2008).